

Z-score-based modularity for community detection in networks

Atsushi Miyauchi* and Yasushi Kawase†

Graduate School of Decision Science and Technology, Tokyo Institute of Technology,
Ookayama 2-12-1, Meguro-ku, Tokyo 152-8552, Japan

(Dated: January 9, 2015)

Identifying community structure in networks is an issue of particular interest in network science. The modularity introduced by Newman and Girvan [Phys. Rev. E **69**, 026113 (2004)] is the most popular quality function for community detection in networks. In this study, we identify a problem in the concept of modularity and suggest a solution to overcome this problem. Specifically, we obtain a new quality function for community detection. We refer to the function as *Z-modularity* because it measures the Z-score of a given division with respect to the fraction of the number of edges within communities. Our theoretical analysis shows that Z-modularity mitigates the resolution limit of the original modularity in certain cases. Computational experiments using both artificial networks and well-known real-world networks demonstrate the validity and reliability of the proposed quality function.

PACS numbers: 89.75.Hc, 02.70.-c

I. INTRODUCTION

Many complex systems can be represented as networks. Analyzing the structure and dynamics of these networks provides meaningful information about the underlying systems. In fact, complex networks have attracted significant attention from diverse fields such as physics, informatics, chemistry, biology, and sociology [1, 2].

An issue of particular interest in network science is the identification of community structure [3]. Roughly speaking, a *community* (also referred to as a *module*) is a subset of vertices more densely connected with each other than with nodes in the rest of the network. Note that no absolute definition of a community exists because any such definition typically depends on the specific system at hand. Detecting communities is a powerful way to discover components that have some special roles or possess important functions. For example, consider the network representing the World Wide Web, where vertices correspond to web pages and edges represent the hyperlinks between pages. Communities in this network are likely to be the sets of web pages dealing with the same or similar topics.

There are various methods to detect community structure in networks, which can be roughly divided into two types. First, there are methods based on some conditions that should be satisfied by a community. The most fundamental concept is a clique. A *clique* is a subset of vertices wherein every pair of vertices is connected by an edge. As even a singleton or an edge is a clique, we are usually interested in finding a *maximum clique* or a *maximal clique*, i.e., cliques with maximum size and cliques not contained in any other clique, respectively. Although the definition of a clique is very intuitive, it is too strong and restrictive to use practically. In 2004, Radicchi *et*

al. [4] introduced more practical definitions: a community in a strong sense and a community in a weak sense. A subset S of vertices is called a *community in a strong sense* if for every vertex in S , the number of neighbors in S is strictly greater than the number of neighbors outside S . On the other hand, a subset S of vertices is called a *community in a weak sense* if the sum, over all vertices in S , of the number of neighbors in S is strictly greater than the number of cut edges of S . Thus, if a subset of vertices is a community in a strong sense, then it is also a community in a weak sense. Recently, Caferi *et al.* [5] proposed an enumerative algorithm to list all divisions of the set of vertices into communities in a strong sense with moderate sizes.

Second, but perhaps more importantly, there are methods that maximize a globally defined quality function. The best known and most commonly used quality function is *modularity*, which was introduced by Newman and Girvan [6]. Here let $G = (V, E)$ be an undirected network consisting of $n = |V|$ vertices and $m = |E|$ edges. The modularity, a quality function for division $\mathcal{C} = \{C_1, \dots, C_k\}$ of V (i.e., $\bigcup_{i=1}^k C_i = V$ and $C_i \cap C_j = \emptyset$ for $i \neq j$), can be written as

$$Q(\mathcal{C}) = \sum_{C \in \mathcal{C}} \left(\frac{m_C}{m} - \left(\frac{D_C}{2m} \right)^2 \right),$$

where m_C is the number of edges in community C , and D_C is the sum of the degrees of the vertices in community C . The modularity represents the sum, over all communities, of the fraction of the number of edges in the communities minus the expected fraction of such edges assuming that they are placed at random with the same distribution of vertex degree.

Many studies have examined modularity maximization. In 2008, Brandes *et al.* [7] proved that modularity maximization is NP-hard. This implies that unless $P = NP$, no modularity maximization method that simultaneously satisfies the following exists: (i) finds a division that

* miyauchi.a.aa@m.titech.ac.jp

† kawase.y.ab@m.titech.ac.jp

maximizes modularity exactly (ii) in time polynomial in n and m (iii) for any networks. To date, a major focus in modularity maximization has been designing accurate and scalable heuristics. In fact, there are a wide variety of algorithms based on greedy techniques [6, 8, 9], simulated annealing [10–12], extremal optimization [13], spectral optimization [14, 15], mathematical programming [16–19], and other techniques. Note that to reduce computation time, a few pre-processing techniques have been proposed [20]. Moreover, to improve the quality of divisions obtained by such heuristics, some post-processing algorithms have also been developed [21].

Although modularity maximization is the most popular and widely used method in practice, it is also known to have some serious drawbacks; i.e., the *resolution limit* [22] and *degeneracies* [23]. The former means that modularity maximization fails to detect communities smaller than a certain scale depending on the total number of edges in a network even if the communities are cliques connected by single edges. The latter means that there exist numerous nearly optimal divisions in terms of modularity maximization, which makes finding communities with maximum modularity extremely difficult. The resolution limit particularly narrows the application range of modularity maximization because most real-world networks consist of communities with very different sizes. To avoid this issue, some multiresolution variants of the modularity have been adopted in practical applications [24–26]. In these variants, the resolution level can be tuned freely by adjusting certain parameters. However, once the resolution level is determined, communities larger than the determined resolution level tend to be divided and smaller communities tend to be merged. Therefore, such multiresolution variants also fail to detect real community structure [27].

In this study, we identify a problem in the concept of modularity and suggest a solution to overcome this problem. Specifically, we obtain a new quality function for community detection. We refer to this function as *Z-modularity* because it measures the Z-score of a given division with respect to the fraction of the number of edges within communities. Our theoretical analysis shows that Z-modularity mitigates the resolution limit of the original modularity in certain cases. In fact, Z-modularity never merges adjacent cliques in the well-known ring of cliques network with any number and size of cliques. Computational experiments using both artificial networks and well-known real-world networks demonstrate the validity and reliability of the proposed quality function.

Note that there are many quality functions based on modularity or other concepts [28–33]. Most of them are collected in Ref. [3].

This paper is structured as follows. In Sec. II, our quality function Z-modularity is introduced. In Sec. III, a theoretical analysis of the properties of Z-modularity is described. The results of computational experiments are shown in Sec. IV. Finally, conclusions and suggestions for future work are given in Sec. V.

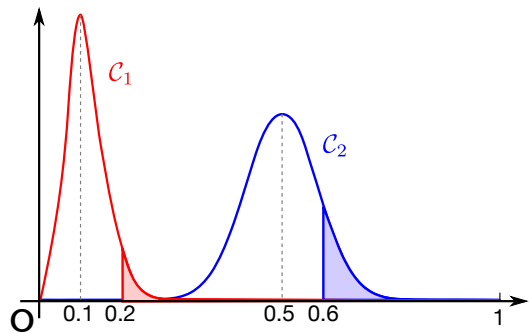


FIG. 1: (Color online) Probability distributions.

II. DEFINITION OF Z-MODULARITY

Modularity simply computes the fraction of the number of edges within communities minus its expected value. The definition is quite intuitive; thus, it is the most popular and widely used quality function in practice.

However, we identify a problem with the concept of modularity. Here consider two divisions \mathcal{C}_1 and \mathcal{C}_2 . Assume that the fraction of the number of edges within communities of \mathcal{C}_1 and \mathcal{C}_2 are 0.2 and 0.6, respectively. In addition, assume that their expected values are 0.1 and 0.5, respectively. Then, we see that these two divisions share the same modularity value (i.e., $Q(\mathcal{C}_1) = Q(\mathcal{C}_2) = 0.1$). The key question is as follows: should these two divisions receive the same quality value? Our answer is that it must depend on the variance of the probability distribution of the fraction of the number of edges within communities of \mathcal{C}_1 and \mathcal{C}_2 . Fig. 1 illustrates an example. In this case, we wish to assign a higher quality value to \mathcal{C}_1 because it is statistically much rarer than \mathcal{C}_2 . This simple but critical observation forms the basis of our quality function.

Given an undirected network $G = (V, E)$ consisting of $n = |V|$ vertices and $m = |E|$ edges, and a division \mathcal{C} of V , we aim to quantify the statistical rarity of division \mathcal{C} in terms of the fraction of the number of edges within communities. To this end, we consider the following edge generation process over V . Place N edges over V at random with the same distribution of vertex degree. Then, when we place an edge, the probability that the edge is placed within communities is given by

$$p = \sum_{C \in \mathcal{C}} \left(\frac{D_C}{2m} \right)^2.$$

Note that this edge generation process is the same as the null-model (also known as the *configuration model* [34]) used in the definition of modularity, with the exception of the sample size. We simply wish to estimate the probability distribution of the fraction of the number of edges within communities. Thus, unlike the null-model, the sample size N is not necessarily equal to the number of edges m .

Let X be a random variable denoting the number of edges generated by the process within communities. Then, X follows the binomial distribution $B(N, p)$. By the central limit theorem, when the sample size N is sufficiently large, the distribution of X/N can be approximated by the normal distribution $\mathcal{N}(p, p(1-p)/N)$. Thus, we can quantify the statistical rarity of division \mathcal{C} in terms of the fraction of the number of edges within communities using the Z-score as follows:

$$Z(\mathcal{C}) = \frac{\sum_{C \in \mathcal{C}} \frac{m_C}{m} - \sum_{C \in \mathcal{C}} \left(\frac{D_C}{2m}\right)^2}{\sqrt{\sum_{C \in \mathcal{C}} \left(\frac{D_C}{2m}\right)^2 \left(1 - \sum_{C \in \mathcal{C}} \left(\frac{D_C}{2m}\right)^2\right)}}.$$

The sample size N never depends on a given division; thus, it is omitted in the denominator. We refer to this quality function as *Z-modularity*.

III. THEORETICAL ANALYSIS

Fortunato and Barthélemy [22] pointed out the resolution limit of modularity. This resolution limit means that modularity maximization fails to detect communities that are smaller than a certain scale depending on the total number of edges in a network. This phenomenon occurs even if the communities are cliques connected by single edges. Here we theoretically analyze Z-modularity from a resolution limit perspective. As a result, we demonstrate that Z-modularity mitigates the resolution limit of the original modularity in certain cases.

A. Ring of cliques network

First, we consider a *ring of cliques network* that consists of a number of cliques connected by single edges (Fig. 2). Assume that each clique consists of p (≥ 3) vertices and the number of cliques is q (≥ 2). Then, the network has $n = p \cdot q$ vertices and $m = q \cdot (1 + p(p-1)/2)$ edges. Fortunato and Barthélemy [22] showed that modularity maximization would merge adjacent cliques if q is larger than a certain value depending on p . However, adjacent cliques are never merged in a division with maximal Z-modularity value, as shown below.

Let \mathcal{C}^* be the division of V into the cliques. In addition, let $\mathcal{C} = \{C_1, \dots, C_l\}$ ($1 < l < q$) be a division of V such that each C_i consists of a series of s_i (≥ 1) cliques and $q = \sum_{i=1}^l s_i$. Then, Z-modularity for \mathcal{C}^* and \mathcal{C} are calculated by

$$Z(\mathcal{C}^*) = \frac{1 - q/m - 1/q}{\sqrt{(1 - 1/q)/q}} \quad \text{and} \quad Z(\mathcal{C}) = \frac{1 - l/m - t}{\sqrt{t(1-t)}},$$

respectively, where $t = \sum_{i=1}^l (s_i/q)^2$. By the Cauchy-Schwarz inequality, we have $1 > t = \sum_{i=1}^l (s_i/q)^2 \geq$

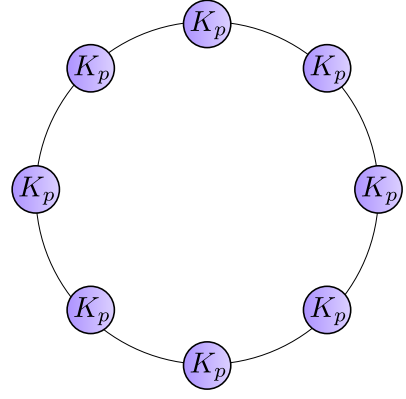


FIG. 2: (Color online) Ring of cliques network. K_p represents a clique with p vertices.

$\left(\sum_{i=1}^l (s_i/q)\right)^2 / l = 1/l$. Here define

$$f(x, y) = \frac{1 - y/m - x}{\sqrt{x(1-x)}}.$$

Then, the derivative of $f(x, y)$ with respect to x is

$$\frac{\partial}{\partial x} f(x, y) = \frac{-x \cdot y/m - (1 - y/m)(1-x)}{2 \cdot (x(1-x))^{3/2}} < 0$$

for $0 < x < 1$ and $1 \leq y \leq m$. Thus, we obtain

$$f(1/l, l) \geq f(t, l).$$

Moreover, the derivative of $f(1/y, y)$ with respect to y is

$$\frac{\partial}{\partial y} f(1/y, y) = \frac{(m - 3y)(y - 1) + y}{2m \cdot (y - 1)^{3/2}} > 0$$

for $1 < y < m/3$. Thus, we have

$$f(1/q, q) > f(1/l, l),$$

since $1 < l < q \leq m/4$ by $m = q \cdot (1 + p(p-1)/2) \geq 4q$. Therefore, we have

$$Z(\mathcal{C}^*) = f(1/q, q) > f(1/l, l) \geq f(t, l) = Z(\mathcal{C}),$$

which means that maximizing Z-modularity never merges adjacent cliques.

Table I lists the values of modularity and Z-modularity of divisions \mathcal{C}^* and \mathcal{C} ($s_i = 2$ for $i = 1, \dots, l$) for some ring of cliques networks. As can be seen, the modularity of \mathcal{C} is greater than that of \mathcal{C}^* when the number of cliques is large, which is consistent with Fortunato and Barthélemy [22]. On the other hand, as we proved above, Z-modularity of \mathcal{C}^* is certainly higher than that of \mathcal{C} for every number of cliques.

B. Network with two pairwise identical cliques

Here we consider a *network with two pairwise identical cliques* that consists of a pair of cliques C_1 and C_2 with

TABLE I: Numerical examples of modularity and Z-modularity for some ring of cliques networks.

n	m	p	q	$Q(\mathcal{C}^*)$	$Q(\mathcal{C})$	$Z(\mathcal{C}^*)$	$Z(\mathcal{C})$
100	220	5	20	0.8591	0.8548	3.942	2.848
200	440	5	40	0.8841	0.9045	5.663	4.150
400	880	5	80	0.8966	0.9295	8.070	5.954
5000	11000	5	1000	0.9081	0.9525	28.73	21.32

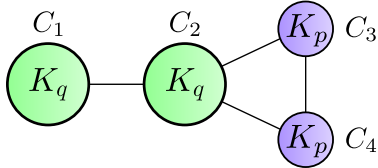


FIG. 3: (Color online) Network with two pairwise identical cliques. K_p and K_q represent cliques with p and q vertices, respectively.

q vertices each and a pair of cliques C_3 and C_4 with p ($< q$) vertices each. These four cliques are connected by single edges, as described in Fig. 3. This network has $n = 2(p + q)$ vertices and $m = p(p - 1) + q(q - 1) + 4$ edges.

Consider two divisions $\mathcal{C}_A = \{C_1, C_2, C_3, C_4\}$ and $\mathcal{C}_B = \{C_1, C_2, C_3 \cup C_4\}$. Note that division \mathcal{C}_A is more natural community structure that we would like to identify. Unfortunately, maximizing Z-modularity may choose \mathcal{C}_B , i.e., $Z(\mathcal{C}_A) < Z(\mathcal{C}_B)$ holds for some pair of p and q . However, if modularity maximization adopts \mathcal{C}_A , then so does Z-modularity, i.e., for any pair of p and q , if $Q(\mathcal{C}_A) > Q(\mathcal{C}_B)$ holds, then $Z(\mathcal{C}_A) > Z(\mathcal{C}_B)$ also holds. This fact follows from the definitions of Z-modularity and the original modularity.

Table II lists the values of modularity and Z-modularity of divisions \mathcal{C}_A and \mathcal{C}_B for some networks with two pairwise identical cliques. We can confirm that both modularity and Z-modularity tend to merge C_3 and C_4 as the sizes of C_1 and C_2 become large. However, there is the case where only Z-modularity could divide C_3 and C_4 . Therefore, we see that Z-modularity again mitigates the resolution limit of modularity in this case.

TABLE II: Numerical examples of modularity and Z-modularity for some networks with two pairwise identical cliques.

n	m	p	q	$Q(\mathcal{C}_A)$	$Q(\mathcal{C}_B)$	$Z(\mathcal{C}_A)$	$Z(\mathcal{C}_B)$
26	80	5	8	0.6618	0.3385	1.443	1.345
42	264	5	16	0.5650	0.5653	1.144	1.143
74	1016	5	32	0.5182	0.5190	1.037	1.039
138	4056	5	64	0.5047	0.5049	1.009	1.010

IV. EXPERIMENTAL RESULTS

The purpose of our computational experiments is to evaluate the validity and reliability of the quality function Z-modularity. To this end, throughout the experiments, we maximize Z-modularity using a simulated annealing algorithm. Note that our algorithm is obtained immediately by changing the objective function from modularity to Z-modularity in the algorithm proposed by Guimerà and Amaral [10]. The implementation of their algorithm can be found on Lancichinetti's web page [35], and we use it with default parameters with the exception of the above change of objective function. Our experiments are conducted on various artificial networks and on well-known real-world networks.

A. Artificial networks

First, we report the results of computational experiments with artificial networks. We compare divisions obtained by maximizing Z-modularity with divisions obtained by modularity maximization on a wide variety of networks. The modularity is also maximized by the simulated annealing algorithm proposed by Guimerà and Amaral [10]. We deal with three types of artificial networks: the planted l -partition model, the Lancichinetti–Fortunato–Radicchi (LFR) benchmark, and the Hanoi graph. For the planted l -partition model and the LFR benchmark, once their parameters are set, the ground-truth community structure is fixed. Thus, we can evaluate the quality of the obtained community structure by comparison with the ground-truth using some measure.

To this end, we adopt the *normalized mutual information* introduced by Danon *et al.* [36]. The normalized mutual information for two divisions \mathcal{C}_1 and \mathcal{C}_2 of n vertices is defined as follows:

$$I_{\text{norm}}(\mathcal{C}_1, \mathcal{C}_2) = \frac{2I(\mathcal{C}_1, \mathcal{C}_2)}{H(\mathcal{C}_1) + H(\mathcal{C}_2)},$$

where

$$I(\mathcal{C}_1, \mathcal{C}_2) = \sum_{C_1 \in \mathcal{C}_1} \sum_{C_2 \in \mathcal{C}_2} \frac{|C_1 \cap C_2|}{n} \log_2 \left(\frac{n \cdot |C_1 \cap C_2|}{|C_1| \cdot |C_2|} \right)$$

and

$$H(\mathcal{C}) = - \sum_{C \in \mathcal{C}} \frac{|C|}{n} \log_2 \frac{|C|}{n}.$$

The normalized mutual information ranges from 0 to 1. For two divisions \mathcal{C}_1 and \mathcal{C}_2 , the higher the normalized mutual information is, the more similar they are (and vice versa). In fact, $I_{\text{norm}}(\mathcal{C}_1, \mathcal{C}_2) = 1$ if \mathcal{C}_1 and \mathcal{C}_2 are identical, and $I_{\text{norm}}(\mathcal{C}_1, \mathcal{C}_2) = 0$ if they are independent. This measure has often been used to evaluate community detection methods. For example, see the computational experiments in Refs. [37, 38].

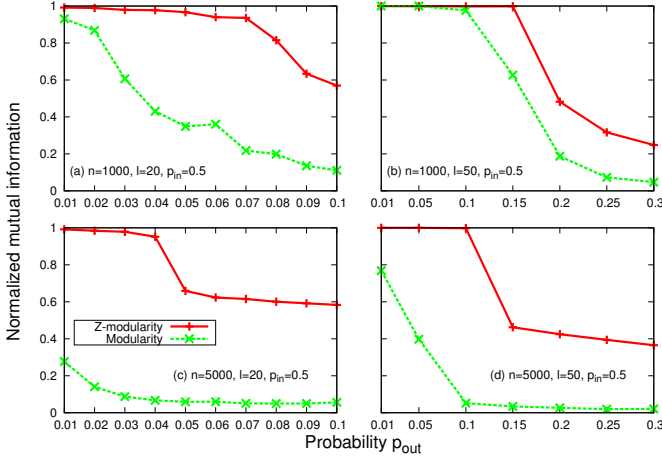


FIG. 4: (Color online) Results for the planted l -partition model.

Planted l -partition model. The planted l -partition model was introduced by Condon and Karp [39]. In this model, n vertices are divided into l equally sized groups. Two vertices in the same group are connected by probability p_{in} , whereas two vertices in different groups are connected by probability $p_{\text{out}} (< p_{\text{in}})$. Throughout the experiments, we set $p_{\text{in}} = 0.5$. We construct four networks corresponding to combinations of two different network sizes ($n = 1000$ or 5000) and two different community sizes ($l = 20$ or 50). The parameter p_{out} starts with 0.01 and then increases in stages.

The results are shown in Fig. 4. As can be seen, Z-modularity outperforms the original modularity in all four cases. In particular, Z-modularity provides much more superior results compared to modularity for networks consisting of relatively small communities.

LFR benchmark. In the planted l -partition model, each group in a generated network forms the Erdős-Rényi random graph [40]. Thus, all vertices have approximately the same degree. Moreover, all groups have exactly the same size. These phenomena are rarely observed in networks in real-world systems. As a more realistic model, the LFR benchmark was proposed by Lancichinetti, Fortunato, and Radicchi [41] for the case of unweighted and undirected networks. The LFR benchmark was then extended to the case of directed and weighted networks with overlapping communities [42]. We now use the original unweighted and undirected case without overlapping communities.

In the model, degree distribution and community size distribution follow the power law with exponents γ and β , respectively. Furthermore, we can specify the number of vertices n , average degree $\langle k \rangle$, maximum degree k_{max} , minimum community size c_{min} , maximum community size c_{max} , and mixing parameter μ . In particular, mixing parameter μ indicates the mixing ratio of communities, i.e., the higher μ is, the more densely connected the communities are. The model constructs a network

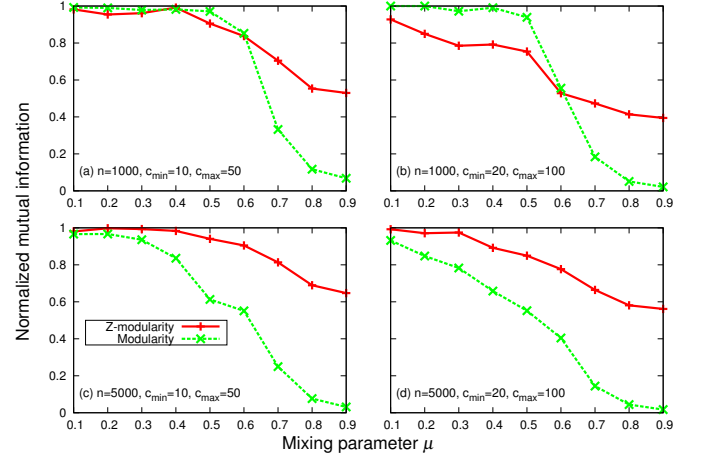


FIG. 5: (Color online) Results for the LFR benchmark.

consistent with the specified parameters. For more details, see Ref. [41]. In our experiments, we set the parameters the same as used in Refs. [37, 38] as follows: $\gamma = -2$, $\beta = -1$, $\langle k \rangle = 20$, and $k_{\text{max}} = 50$. We construct four networks corresponding to combinations of two different network sizes ($n = 1000$ or 5000) and two different ranges of community size ($(c_{\text{min}}, c_{\text{max}}) = (10, 50)$ or $(20, 100)$).

The results are illustrated in Fig. 5. For the smaller networks ($n = 1000$), the mutual information values obtained by maximizing Z-modularity are lower than those obtained by modularity maximization when $\mu \leq 0.6$ for both community size settings. This trend is significant when the network consists of relatively large communities ($(c_{\text{min}}, c_{\text{max}}) = (20, 100)$). On the other hand, for larger networks ($n = 5000$), Z-modularity outperforms the original modularity for both community size settings. From the above, we see that Z-modularity is particularly suitable for identifying community structure when a network consists of relatively small communities.

Here we investigate why the mutual information values obtained by maximizing Z-modularity are low when the community sizes are large. To this end, Fig. 6 depicts the adjacency matrices of the LFR benchmark network with parameters $\gamma = -2$, $\beta = -1$, $n = 1000$, $\langle k \rangle = 20$, $k_{\text{max}} = 50$, $c_{\text{min}} = 20$, $c_{\text{max}} = 100$, and $\mu = 0.3$. The vertices are ordered according to both the ground-truth partition and the optimal partition for Z-modularity. The edges connecting vertices in the same community and in different communities are plotted with different colors, i.e., red and blue, respectively. As can be seen, maximizing Z-modularity divides the relatively large ground-truth communities because they contain much denser communities in the hierarchical structure by random behavior.

Hanoi graph. Here we demonstrate optimal partitions with respect to Z-modularity and the original modularity for the Hanoi graph, which is an example of networks with hierarchical organization. The Hanoi graph H_n con-

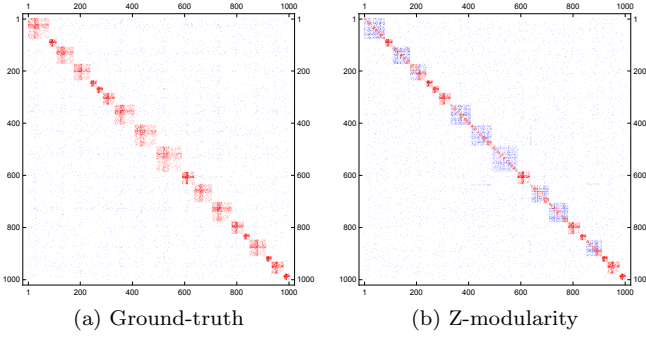


FIG. 6: (Color online) Adjacency matrices for an LFR benchmark network.

responds to the allowed moves in the *tower of Hanoi* for n disks, which is a famous puzzle invented by Édouard Lucas in 1883. The Hanoi graph H_n has 3^n vertices and $3 \cdot (3^n - 1)/2$ edges. In the context of community detection in networks, the Hanoi graph H_3 is used by Rosvall and Bergstrom [43].

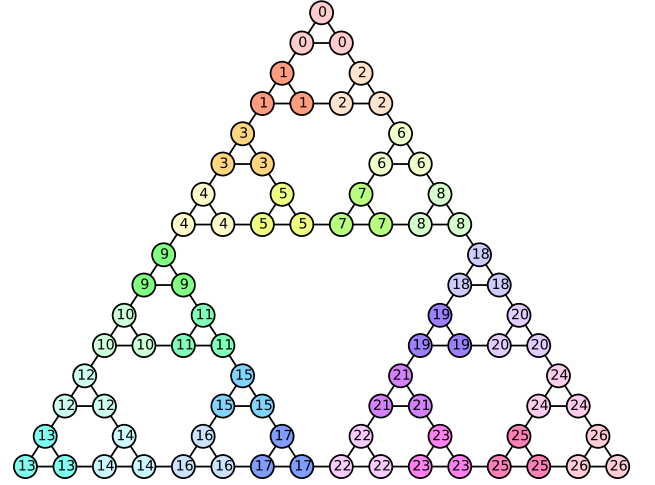
The results for Hanoi graph H_4 are shown in Fig. 7, where the label (and color) of each vertex represents the community to which the vertex belongs. As can be seen, maximizing Z-modularity leads to more detailed partition than modularity maximization.

B. Real-world networks

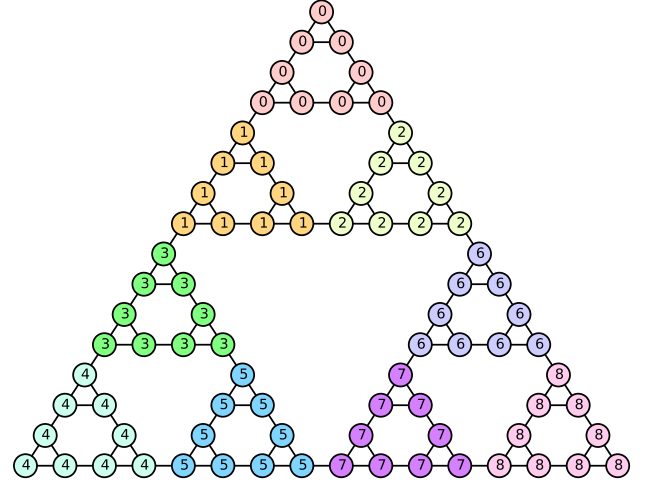
Here we report the results of computational experiments with real-world networks; i.e., the Zachary's karate club network, the Les Misérables network, and the American college football network.

Zachary's karate club network. The first example is the famous karate club network analyzed by Zachary [44], which is often used as a benchmark to evaluate community detection methods. It consists of 34 vertices representing the members in a karate club in an American university, in addition to 78 edges representing friendship relations among individuals. Because of a conflict between the club administrator and the instructor, the club members split into two groups, one supporting the administrator and the other supporting the instructor. Therefore, these groups can be viewed as a ground-truth community structure.

The division obtained by maximizing Z-modularity is shown in Fig. 8, where vertices with the same color represent a community. The label of each vertex represents an identification number of the member. For example, 1 and 34 represent the administrator and the instructor, respectively. The dashed line gives the division of the network into the above two groups. Although the community $\{3, 10, 29\}$ straddles two groups, the other communities are all contained in either one of the groups.



(a) Optimal partition for Z-modularity: 27 communities, $Z = 3.376$, and $Q = 0.6379$.



(b) Optimal partition for modularity: 9 communities, $Z = 2.510$, and $Q = 0.7889$.

FIG. 7: (Color online) Community structure for Hanoi graph H_4 .

Les Misérables network. The second example is the network of the characters in the novel *Les Misérables* by Victor Hugo, compiled by Knuth [45]. It consists of 77 vertices representing the characters and 254 edges indicating the co-appearance of characters.

The division obtained by maximizing Z-modularity is presented in Fig. 9, where vertices with the same color represent a community. The label of each vertex represents the name of the character. Identified communities are likely to correspond to specific groups within the story. For example, the community consisting of 12 vertices (shaded with light brown) at the top left corner contains major characters belonging to the revolutionary student club *Friends of the ABC*.

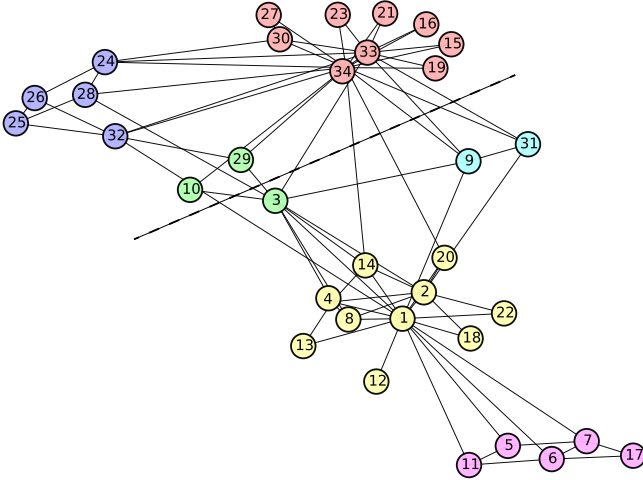


FIG. 8: (Color online) Community structure for Zachary's karate club network: 6 communities, $Z = 0.9266$, and $Q = 0.3882$.

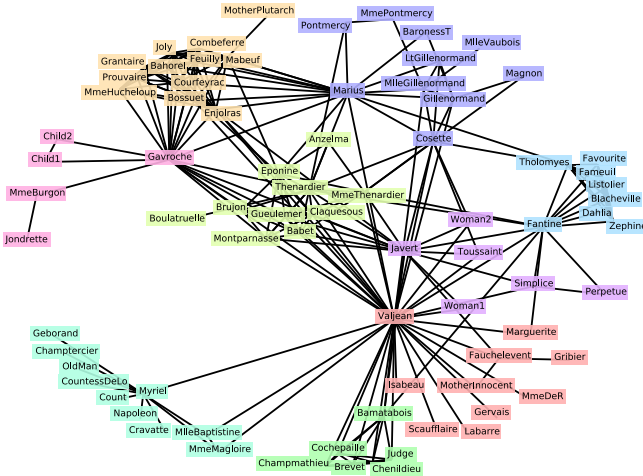


FIG. 9: (Color online) Community structure for Les Misérables network: 9 communities, $Z = 1.490$, and $Q = 0.5245$.

American college football network. The third and final example is a network of college football teams in the United States, which was derived by Girvan and Newman [46]. There are 115 vertices representing the football teams, and 654 edges connecting teams that played each other in a regular season. The teams are divided into 12 groups referred to as *conferences* containing approximately 10 teams each. More games are played between teams in the same conference than between teams in different conferences. Thus, the conferences can be viewed as a ground-truth community structure.

The division obtained by maximizing Z-modularity is shown in Fig. 10, where vertices with the same color represent a community. Note that the label of each vertex now represents the conference to which the team

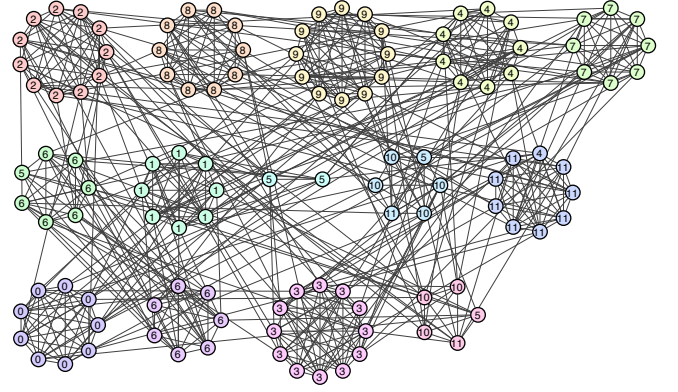


FIG. 10: (Color online) Community structure for American college football network: 14 communities, $Z = 2.111$, and $Q = 0.5738$.

belongs rather than an identification number of the team. Although some misclassifications are observed, Z-modularity correctly identifies 7 out of 12 conferences (i.e., conferences 0, 1, 2, 3, 7, 8, and 9). This result is outstanding in comparison with divisions obtained by modularity maximization. In fact, as reported in Ref. [16], only four conferences were correctly recovered by division with a higher modularity value $Q = 0.6046$.

V. CONCLUSIONS

In this study, we have identified a problem in the concept of modularity and suggested a solution to overcome this problem. Specifically, we have obtained a new quality function Z-modularity that measures the Z-score of a given division with respect to the fraction of the number of edges within communities. Theoretical analysis has shown that Z-modularity mitigates the resolution limit of the original modularity in certain cases. In fact, Z-modularity never merges adjacent cliques in the well-known ring of cliques network with any number and size of cliques. In computational experiments, we have evaluated the validity and reliability of Z-modularity. The results for artificial networks show that Z-modularity more accurately detects the ground-truth community structure than the original modularity in most cases. In particular, Z-modularity outperforms modularity for networks consisting of relatively small communities. Furthermore, the results for real-world networks demonstrate that Z-modularity leads to natural and reasonable community structure in practical use. Therefore, we conclude that Z-modularity could be another option for the quality function in community detection.

In the future, further experiments should be conducted to examine the performance of Z-modularity in more details. Although strict experiments were conducted in the present study, other experimental settings are also possible. As another future direction, the physical interpreta-

tion of maximizing Z-modularity should be investigated. For example, it is known that modularity maximization can be interpreted as the problem of finding the ground state of a spin glass model [24].

ACKNOWLEDGMENTS

The first author is supported by a Grant-in-Aid for JSPS Fellows (No. 26-11908). The second author is supported by a Grant-in-Aid for Research Activity Start-up (No. 26887014). This work was partially supported by JST, ERATO, Kawarabayashi Large Graph Project.

-
- [1] M. E. J. Newman, *SIAM Rev.* **45**, 167 (2003).
 - [2] M. E. J. Newman, *Networks: An Introduction* (Oxford University Press, 2010).
 - [3] S. Fortunato, *Phys. Rep.* **486**, 75 (2010).
 - [4] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 2658 (2004).
 - [5] S. Cafieri, G. Caporossi, P. Hansen, S. Perron, and A. Costa, *Phys. Rev. E* **85**, 046113 (2012).
 - [6] M. E. J. Newman and M. Girvan, *Phys. Rev. E* **69**, 026113 (2004).
 - [7] U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hofer, Z. Nikoloski, and D. Wagner, *IEEE Trans. Knowl. Data Eng.* **20**, 172 (2008).
 - [8] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, *J. Stat. Mech.: Theory Exp.* (2008) P10008.
 - [9] A. Clauset, M. E. J. Newman, and C. Moore, *Phys. Rev. E* **70**, 066111 (2004).
 - [10] R. Guimerà and L. A. N. Amaral, *Nature (London)* **433**, 895 (2005).
 - [11] C. P. Massen and J. P. K. Doye, *Phys. Rev. E* **71**, 046101 (2005).
 - [12] A. Medus, G. Acuña, and C. Dorso, *Physica A* **358**, 593 (2005).
 - [13] J. Duch and A. Arenas, *Phys. Rev. E* **72**, 027104 (2005).
 - [14] M. E. J. Newman, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 8577 (2006).
 - [15] T. Richardson, P. J. Mucha, and M. A. Porter, *Phys. Rev. E* **80**, 036111 (2009).
 - [16] G. Agarwal and D. Kempe, *Eur. Phys. J. B* **66**, 409 (2008).
 - [17] S. Cafieri, P. Hansen, and L. Liberti, *Phys. Rev. E* **83**, 056105 (2011).
 - [18] A. Miyauchi and Y. Miyamoto, *Eur. Phys. J. B* **86**, 302 (2013).
 - [19] S. Cafieri, A. Costa, and P. Hansen, *Ann. Oper. Res.* **222**, 213 (2014).
 - [20] A. Arenas, J. Duch, A. Fernández, and S. Gómez, *New J. Phys.* **9**, 176 (2007).
 - [21] S. Cafieri, P. Hansen, and L. Liberti, *Discr. Appl. Math.* **163**, 65 (2014).
 - [22] S. Fortunato and M. Barthélemy, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 36 (2007).
 - [23] B. H. Good, Y.-A. de Montjoye, and A. Clauset, *Phys. Rev. E* **81**, 046106 (2010).
 - [24] J. Reichardt and S. Bornholdt, *Phys. Rev. E* **74**, 016110 (2006).
 - [25] A. Arenas, A. Fernández, and S. Gómez, *New J. Phys.* **10**, 053039 (2008).
 - [26] P. Pons and M. Latapy, *Theor. Comput. Sci.* **412**, 892 (2011).
 - [27] A. Lancichinetti and S. Fortunato, *Phys. Rev. E* **84**, 066122 (2011).
 - [28] M. Rosvall and C. T. Bergstrom, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 7327 (2007).
 - [29] Z. Li, S. Zhang, R.-S. Wang, X.-S. Zhang, and L. Chen, *Phys. Rev. E* **77**, 036109 (2008).
 - [30] M. Rosvall and C. T. Bergstrom, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 1118 (2008).
 - [31] J. Chen, O. R. Zaïane, and R. Goebel, in *Proceedings of the 2009 SIAM International Conference on Data Mining*, Vol. 3, edited by C. Apte, H. Park, K. Wang, and M. J. Zaki (SIAM, 2009) pp. 978–989.
 - [32] S. Zhang and H. Zhao, *Phys. Rev. E* **85**, 066114 (2012).
 - [33] S. Zhang and H. Zhao, *Phys. Rev. E* **88**, 052802 (2013).
 - [34] M. Molloy and B. Reed, *Random Struct. Algorithms* **6**, 161 (1995).
 - [35] A. Lancichinetti, <https://sites.google.com/site/andrealancichinetti/software>.
 - [36] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas, *J. Stat. Mech.: Theory Exp.* (2005) P09008.
 - [37] A. Lancichinetti and S. Fortunato, *Phys. Rev. E* **80**, 056117 (2009).
 - [38] A. Lancichinetti and S. Fortunato, *Phys. Rev. E* **89**, 049902 (2014).
 - [39] A. Condon and R. M. Karp, *Random Struct. Algorithms* **18**, 116 (2001).
 - [40] P. Erdős and A. Rényi, *Publ. Math. Debrecen* **6**, 290 (1959).
 - [41] A. Lancichinetti, S. Fortunato, and F. Radicchi, *Phys. Rev. E* **78**, 046110 (2008).
 - [42] A. Lancichinetti and S. Fortunato, *Phys. Rev. E* **80**, 016118 (2009).
 - [43] M. Rosvall and C. T. Bergstrom, *PLoS ONE* **6**, e18209 (2011).
 - [44] W. W. Zachary, *J. Anthropol. Res.* **33**, 452 (1977).
 - [45] D. E. Knuth, *The Stanford GraphBase: A Platform for Combinatorial Computing* (Addison-Wesley, Reading, MA, 1993).
 - [46] M. Girvan and M. E. J. Newman, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 7821 (2002).